AT&T Technical Journal
Vol. 64, No. 6, July-August 1985
Printed in U.S.A.

Some Properties of Continuous Hidden Markov Model Representations

By L. R. RABINER, B.-H. JUANG, S. E. LEVINSON, and M. M. SONDHI*

(Manuscript received January 14, 1985)

Many signals can be modeled as probabilistic functions of Markov chains in which the observed signal is a random vector whose probability density function (pdf) depends on the current state of an underlying Markov chain. Such models are called Hidden Markov Models (HMMs) and are useful representations for speech signals in terms of some convenient observations (e.g., cepstral coefficients or pseudolog area ratios). One method of estimating parameters of HMMs is the well-known Baum-Welch reestimation method. For continuous pdf's, the method was known to work only for elliptically symmetric densities. We have recently shown that the method can be generalized to handle mixtures of elliptically symmetric pdf's. Any continuous pdf can be approximated to any desired accuracy by such mixtures, in particular, by mixtures of multivariate Gaussian pdf's. To effectively make use of this method of parameter estimation, it is necessary to understand how it is affected by the amount of training data available, the number of states in the Markov chain, the dimensionality of the signal, etc. To study these issues, Markov chains and random vector generators were simulated to generate training sequences from "toy" models. The model parameters were estimated from these training sequences and compared to the "true" parameters by means of an appropriate distance measure. The results of several such experiments show the strong sensitivity of the method to some (but not all) of the model parameters. A procedure for getting good initial parameter estimates is, therefore, of considerable importance.

* Authors are employees of AT&T Bell Laboratories.

I. INTRODUCTION

The theory of signal representation based on Hidden Markov Models (HMMs) is well established and has been applied to text analysis,[1] coding theory,[2] ecology,[3] and most recently, speech processing.[4-6] The form of the HMM that we are considering is sketched in Fig. 1. The Markov chain has N states, and transitions between states are governed by a stochastic transition matrix, $\mathbf{A}$, with elements a_{ij}, where

$$\begin{aligned} a_{ij} = {} & \text{probability of making a transition to state } j, \\ & \text{given currently in state } i. \end{aligned}$$

In a given state, j, the observed output of the model is a random vector with a probability density function (pdf) b_j.

Given the model of Fig. 1, it is necessary to be able to estimate the model parameters (i.e., the transition matrix, $\mathbf{A}$, and the pdf's b_j) from training data consisting of observations of output sequences generated by the model. One very useful method of parameter estimation for HMMs is the Baum-Welch reestimation procedure.[7] For the case of continuous pdf's of interest here, the method was originally shown to be valid for log-concave densities.[7] This restriction was relaxed by Liporace,[8] who extended the applicability of the method to elliptically symmetric densities. However, this class of densities is still too restrictive for many interesting problems (e.g., measured densities of various

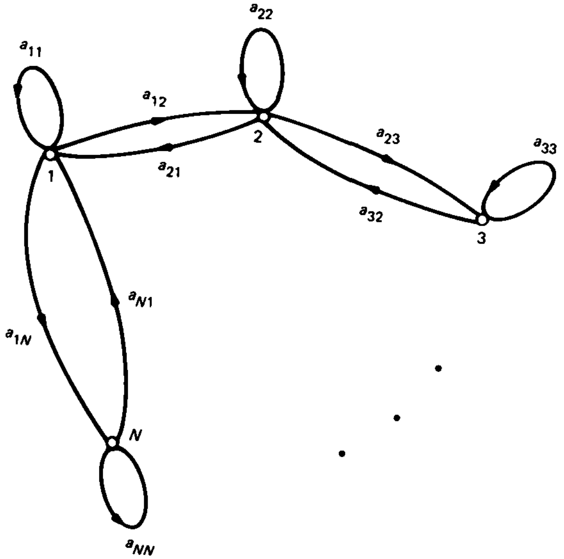

Fig. 1—Markov model with N states.

speech parameters). Therefore, we consider a more general representation of the density—a finite mixture of the form

$$b_j(\mathbf{x}) = \sum_{m=1}^{M} c_{jm}\mathcal{N}[\mathbf{x}, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm}] \qquad j = 1, N, \tag{1}$$

where $\mathcal{N}$ may be any log-concave or elliptically symmetric density, and is assumed to be Gaussian in our present study. The vector $\mathbf{x}$ is the observation vector. The vector $\boldsymbol{\mu}_{jm}$ and the matrix $\mathbf{U}_{jm}$ are, respectively, the mean vector and the covariance matrix for the mth mixture component in state j. The coefficients, c_{jm}, are the mixture gains, and satisfy the stochastic constraint

$$\sum_{m=1}^{M} c_{jm} = 1, \qquad c_{jm} \geq 0, \tag{2}$$

so that

$$\int_{-\infty}^{\infty} b_j(\mathbf{x})d\mathbf{x} = 1, \qquad j = 1, N. \tag{3}$$

The representation of eq. (1) can be used to approximate arbitrarily closely any finite, continuous density function; hence its appropriateness to a wide range of problems. It has recently been shown[9] that the reestimation procedure of Refs. 7 and 8 can be extended to cover the mixture representation of eq. (1).

To understand the properties of such HMMs, and to study the sensitivity of the parameter estimates to the details of the estimation procedure, we have simulated several "toy" models and examined the effects of sample size, initial parameter estimates, model inconsistencies, etc., on the corresponding estimated models. In this paper we present the results of our simulations. Since we have studied only a few, carefully selected cases, we make no claims about specific sample sizes, range of initial parameter values, etc. Instead, it is intended that the examples presented allow the reader to understand the nature of the representation, and thereby use it appropriately for his or her particular application.

The outline of this paper is as follows. In Section I we show how a toy model or HMM sequence generator can be implemented to provide appropriate training sequences for estimating model parameters. In Section II we review the continuous density HMM. In Section III we describe a series of experiments designed to study the sensitivities and properties of the HMM signal representation. Finally, in Section IV we review the key results and discuss their implications for practical problems.

II. REVIEW OF THE CONTINUOUS HMM

Consider an N-state Markov chain where we label the states q_1, q_2, $\cdots$, q_N. The Markov chain is characterized by its state transition matrix $\mathbf{A} = [a_{ij}]$. Each state q_j is characterized by a continuous multivariate, probability density function $b_j(\mathbf{x})$, where $\mathbf{x}$ is a K-dimensional observation vector.

Given a sequence of observations, $\mathbf{O} = O_1, O_2, \cdots, O_T$, where each O_t is a K-dimensional vector, we can calculate the likelihood of $\mathbf{O}$, given a model $\mathbf{M}$. We denote the likelihood as $\mathscr{L}(\mathbf{O} \mid \mathbf{M})$. Following Baum,[10] we can define a set of forward and backward likelihoods, $a_t(i)$ and $\beta_t(j)$ respectively, where, for $1 \le i, j \le N$, and $1 \le t \le T$,

$$\alpha_t(i) = \mathscr{L}(O_1, O_2, \cdots, O_t \text{ and } q_i \text{ at time } t \mid \mathbf{M}) \tag{4}$$

and

$$\beta_t(j) = \mathscr{L}(O_{t+1} O_{t+2} \cdots O_T \mid q_j \text{ at time } t \text{ and } \mathbf{M}). \tag{5}$$

Baum has shown that $\alpha_t(i)$ and $\beta_t(j)$ can be computed recursively. Assuming that we start in q_1, whereby $\alpha_0(1) = 1$, $\alpha_0\ (i) = 0$, $2 \le i \le N$, and $\beta_T(j) = 1$, $1 \le j \le N$, then for $1 \le t \le T$ we get

$$\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} \right] b_j(O_t), \tag{6}$$

and for $T - 1 \ge t \ge 0$,

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j). \tag{7}$$

Thus, $\mathscr{L}(\mathbf{O} \mid \mathbf{M})$ can be efficiently evaluated as

$$\mathscr{L}(\mathbf{O} \mid \mathbf{M}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \tag{8}$$

for $0 \le t \le T - 1$. The parameters of the HMM are estimated by finding some $\mathbf{M}$ that is a local maximum of $\mathscr{L}(\mathbf{O} \mid \mathbf{M})$ for a given observation sequence $\mathbf{O}$.

Using the mixture density of eq. (1) as the parameterized pdf's $b_j(\mathbf{x})$, the model $\mathbf{M}$ is specified by the following:

N = number of states in the model
M = number of mixture densities for each distribution
K = number of dimensions of each observation vector
$\mathbf{A}$ = $[a_{ij}]$ = state transition matrix
$\mathbf{C}$ = $[c_{jm}]$, where c_{jm} = mixture gains for mth mixture in state j
$\boldsymbol{\mu}$ = $[\mu_{jmk}]$, where μ_{jmk} = the kth component of the mean vector μ_{jm} for mth mixture in state j

$\mathbf{U} = [U_{jmkl}]$, where U_{jmkl} = the (k, l)th entry of the covariance matrix for the mth mixture in state j.

Given the values chosen for N, M, and K, and a set of initial guesses for $\mathbf{A}$, $\mathbf{C}$, $\boldsymbol{\mu}$, and $\mathbf{U}$, a set of reestimation formulas is available[9] for optimizing $\mathscr{L}(\mathbf{O} \mid \mathbf{M})$, for a given training set of observations $\mathbf{O}$.

There are two general cases of the model that are of interest, namely the *ergodic* case in which the Markov chain is ergodic (i.e., all states are aperiodic and recurrent nonnull) and the *left-to-right* case in which a transition from state q_i to state q_j is possible only if $j \geq i$ (i.e., there is a sequential progression through states of the model). Both general cases are of interest for real-world applications.

2.1 Toy Markov model generator

In order to investigate the behavior of the parameter estimation algorithms for the continuous HMM, a toy Markov model generator was implemented. Its function was to generate an observation sequence (for the ergodic case), or a set of observation sequences (for the left-to-right case), for an input model specification $\mathbf{M}_{\text{in}}$. Each observation generated by the model was a K-dimensional vector according to the probability density $b_j(\mathbf{x})$ for the jth state.

The algorithm used to generate the observation sequences is the following:

1. Set the state index, $j = 1$ and the time index, $t = 1$.

2. Partition the unit interval proportionally to c_{jm}, $1 \leq m \leq M$. Generate x, a random number uniform on $[0, 1]$. Select the mixture density, l, according to the subinterval in which x falls.

3. Decompose $\mathbf{U}_{jl}$ into $\mathbf{Q\Lambda Q}'$, where $\mathbf{Q}$ is the matrix of eigenvectors of $\mathbf{U}_{jl}$ and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{U}_{jl}$.

4. Generate a K-dimensional normal deviate, $\mathbf{y}$, of zero mean and covariance $\boldsymbol{\Lambda}$.

5. Set $\mathbf{O}_t = \mathbf{Qy} + \boldsymbol{\mu}_{jl}$.

6. Partition the unit interval proportionally to a_{jk}, $1 \leq k \leq N$. Generate x, a random deviate uniform on $[0, 1]$ and select the next state, i, according to the subinterval in which x falls.

7. Increment t.

8. If $t \leq T$ go to 2; else, stop.

The Markov model generator was specified by a model $\mathbf{M}_{\text{in}}$, and by a limit on the number of observations T, or on the number of sequences Q (for the left-to-right case). Each individual sequence, in the left-to-right case, started in state q_1 (at observation 1) and terminated in state q_N (at observation T), with the property that it had to have been in state q_N for at least L observations. (Typically L was 5 to 10.)

III. EXPERIMENTAL EVALUATION OF THE REESTIMATION PROCEDURE

A series of experimental evaluations of the reestimation procedure were made to determine the sensitivities of the algorithm—and hence the resulting HMMs—to aspects of the observation sequence used to train the model. Using the toy Markov model generators, several input source models were defined (i.e., the model parameters were specified) and several sets of observation sequences were generated from the models. For each input source model, the reestimation algorithm was used to obtain locally optimal model parameters based on the generated sequences and initial estimates. The resulting model $\mathbf{M}$ was compared with the source model using a probabilistic distance measure[11] of the form

$$D(\mathbf{M}_{\rm in}, \mathbf{M}) = \frac{\log[\mathscr{L}(\mathbf{O}_{M_{\rm in}} | \mathbf{M}_{\rm in})] - \log[\mathscr{L}(\mathbf{O}_{M_{\rm in}} | \mathbf{M})]}{T_{\rm in}}, \tag{9}$$

where $\mathbf{O}_{M_{\rm in}}$ was a set of observations generated by the toy model $\mathbf{M}_{\rm in}$, and $T_{\rm in}$ was the total number of observations in this set. The distance measure of eq. (9) gives the normalized difference in log likelihoods of the observation sequence coming from the true toy model, and of the likelihood of its coming from the estimated model, where the normalization is the number of observations in $\mathbf{O}_{M_{\rm in}}$. Previous experience with D has shown that this measure is very effective for comparing HMMs.[11]

3.1 Correlation of model distance to changes in model parameters

Before investigating the sensitivities of the reestimation procedure to various model parameters and initial conditions, a preliminary experiment was run to measure the correlation of model distance to changes in model parameters. For this experiment, the initial (ergodic) model had the specifications

$$M = N = K = 2$$

$$\mathbf{A} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0.75 & 0.25 \\ 0.35 & 0.65 \end{bmatrix}$$

$$\boldsymbol{\mu}_{1..} = \begin{bmatrix} 1. & 5. \\ 3. & 4 \end{bmatrix}, \quad \boldsymbol{\mu}_{2..} = \begin{bmatrix} 5. & 9. \\ 8. & 2. \end{bmatrix}$$

$$\mathbf{U}_{11..} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}, \quad \mathbf{U}_{21..} = \begin{bmatrix} 7 & 3 \\ 3 & 7 \end{bmatrix}, \quad \mathbf{U}_{12..} = \begin{bmatrix} 10 & 2 \\ 2 & 10 \end{bmatrix}$$

$$\mathbf{U}_{22..} = \begin{bmatrix} 8 & 1 \\ 1 & 8 \end{bmatrix}.$$

A new model was created in which all model parameters remained the same except for one set, in which the columns of the corresponding matrix or matrices were reversed, i.e., a_{ij} became a_{ji}, or c_{jm} became c_{mj}, etc. In this manner we could study the effects of changing only a single parameter set on the model distance. A smooth interpolation between the parameter set for the initial model and the reversed parameter set was made by changing the parameter set in steps, and then measuring model distance at each step. In particular, if we denote the matrix in the initial model by **X** and the reversed matrix in the new model by **X**$'$, the intermediate matrices **X**$''$ were formed by

$$\mathbf{X}'' = \frac{1}{1+\delta}\mathbf{X} + \frac{\delta}{1+\delta}\mathbf{X}',$$

where the deviation factor $\delta = \epsilon, 2\epsilon, \cdots, 2^{12}\epsilon$ and $\epsilon = 0.016$.

The results of this preliminary experiment are shown in Fig. 2, which gives a series of plots of model distance D, versus signal-to-noise ratio γ, defined as

$$\gamma = 10 \log_{10} \frac{\|\mathbf{X}\|^2}{\|\mathbf{X} - \mathbf{X}''\|^2},$$

where $\|\cdot\|$ denotes matrix norm ($\|\mathbf{X}\|^2 = \sum_i \sum_j x_{ij}^2$ for $\mathbf{X} = [x_{ij}]$) for changes in **A**, **C**, and μ. (Curves similar to that for μ can be obtained for changes in **U**.) It can be seen that perturbed models are far more

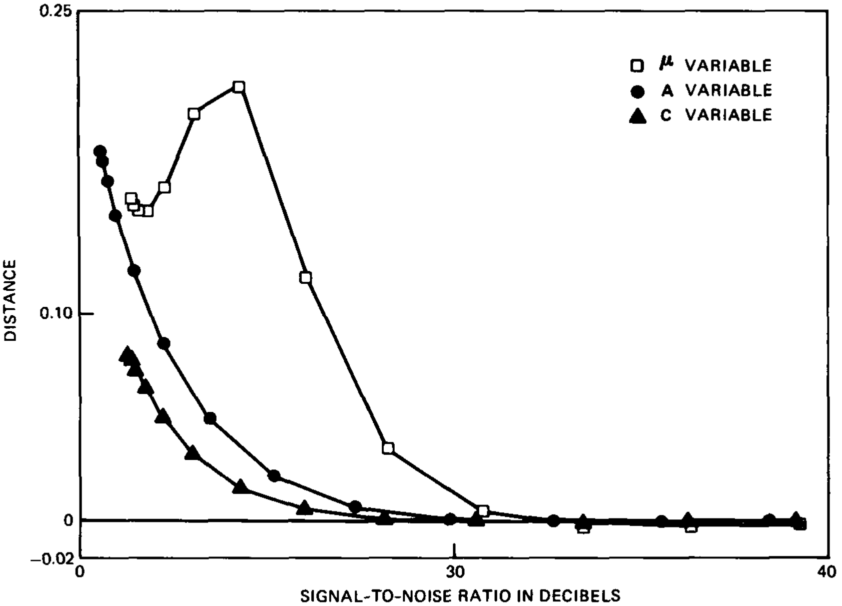

Fig. 2—Distance as a function of parameter deviation for changes in μ, **A**, and **C**.

distant (in the probabilistic distance sense) from the initial models when μ is perturbed than when the transition probability or the mixture gain parameters are perturbed. In general, the HMMs are indeed much more sensitive to small errors in μ values than to small errors in **C** or **A** values unless the variances are extremely large. The exact sensitivity will depend on the precise relationships among means and the associated variances.

The effects of the detailed relationships among means and the associated variances can be seen from the nonmonotonic behavior of the distance curve pertaining to μ in Fig. 2. In this particular case, the mean vectors moved from (1,5) to (5,1), (3,4) to (4,3), (5,9) to (9,5), and (8,2) to (2,8) as the signal-to-noise ratio decreased from about 38.6 dB to 2.6 dB. Thus, as seen in Fig. 2, when γ drops below 8 dB, the probabilistic distance for μ deviations actually decreases. One may observe that in some section along the perturbation path from (8,2) to (2,8), the perturbed mean moves *closer* to the original mean locations (1,5) and (3,4), and thus results in a decrease rather than increase in the probabilistic distance.

We should point out that if we arbitrarily increase the number of mixture components in modeling a given density, then, with proper choice of the initial estimate, the obtained mixture weights become proportional to sample values of the density function at the mean locations of the mixtures. When this happens, the variance in each mixture density, as well as the spacing of the means, decreases and, asymptotically, the observation density is mainly characterized by the mixture gains and the mean vectors. Thus, there is a continuum in the observation of relative model sensitivities as the number of mixture terms varies.

3.2 Sensitivities of the reestimation procedure to parameter inaccuracies and to the training sequence

Based on the results given in the previous section, a series of experiments were performed to investigate the sensitivities of the reestimation procedure to initial parameter estimates and to the length of the training sequence.

The first experiment used a left-to-right source model with the characteristics

$$N = 5, \qquad M = 3, \qquad K = 5$$

$$\mathbf{A} = \begin{bmatrix} .8 & .15 & .05 & 0 & 0 \\ 0 & .8 & .15 & .05 & 0 \\ 0 & 0 & .8 & .15 & .05 \\ 0 & 0 & 0 & .8 & .2 \\ 0 & 0 & 0 & 0 & 1. \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} .6 & .3 & .1 \\ .6 & .3 & .1 \\ .6 & .3 & .1 \\ .6 & .3 & .1 \\ .6 & .3 & .1 \\ .6 & .3 & .1 \end{bmatrix}$$

$$\boldsymbol{\mu}_{1..} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1.5 & 1.5 & 1.5 & 1.5 & 1.5 \\ 2 & 2 & 2 & 2 & 2 \end{bmatrix}, \quad \boldsymbol{\mu}_{2..} = \begin{bmatrix} 3 & 3 & 3 & 3 & 3 \\ 3.5 & 3.5 & 3.5 & 3.5 & 3.5 \\ 4 & 4 & 4 & 4 & 4 \end{bmatrix},$$

$$\boldsymbol{\mu}_{3..} = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 5.5 & 5.5 & 5.5 & 5.5 & 5.5 \\ 6 & 6 & 6 & 6 & 6 \end{bmatrix}$$

$$\boldsymbol{\mu}_{4..} = \begin{bmatrix} 7 & 7 & 7 & 7 & 7 \\ 7.5 & 7.5 & 7.5 & 7.5 & 7.5 \\ 8 & 8 & 8 & 8 & 8 \end{bmatrix}, \quad \boldsymbol{\mu}_{5..} = \begin{bmatrix} 9 & 9 & 9 & 9 & 9 \\ 9.5 & 9.5 & 9.5 & 9.5 & 9.5 \\ 10 & 10 & 10 & 10 & 10 \end{bmatrix}$$

$$\mathbf{U}_{jmkk} = \begin{cases} .5 & m = 1 \\ .2 & m = 2 \\ .1 & m = 3 \end{cases} \quad \text{for} \quad k = 1, 2, \cdots, 5, \quad \text{and}$$

$$\mathbf{U}_{jmkl} = 0.01 \quad \text{for all} \quad k \neq l.$$

The initial guess of the model parameters was random for **A** and **C**, and identity matrices for **U**. For $\boldsymbol{\mu}$, the initial guess had the form

$$\boldsymbol{\mu}' = (1 - z \cdot \alpha)\boldsymbol{\mu}, \tag{10}$$

where α is a random variable uniformly distributed on (0,2), and z is a user-specified error bound, which limits the maximum possible deviation of $\boldsymbol{\mu}'$ from $\boldsymbol{\mu}$ in the source model.

The source generated Q random sequences according to the specified model, where Q varied from 10 to 100 (in steps of 10) and initial estimates with values of $z = 0.0$, 0.2, 0.4, and 0.6 were used. For each set of observations, and for each initial estimate, the reestimation procedure was iterated until a stationary point was found. At this point, both the average (negative) log likelihood for the estimated model **M** and the model distance (from the source to the estimated model) were calculated. Figure 3 shows a series of plots of the average (negative) log likelihood, and the model distance, as a function of the total number of observations in the Q training sequences, for the four values of z. For values of $z = 0$ and 0.2, for sufficiently long training sequences (i.e., 20 sets of observations or about 400 observations), the model distances were reasonably small (less than 0.25). As z got bigger, thereby making the initial estimates of $\boldsymbol{\mu}$ poorer, the resulting models had distances on the order of 0.4 or larger. It is clearly shown that the accuracy of the estimated model depends on the initial estimate from which the iterative reestimation procedure starts. A converged model estimate is only a local optimum and, in general, has a lower likelihood than the global optimum.

Figure 3 also shows the correlation between the log likelihood and the model distance. For sufficiently long observations, the model distance is a good predictor of the relative log likelihood for the

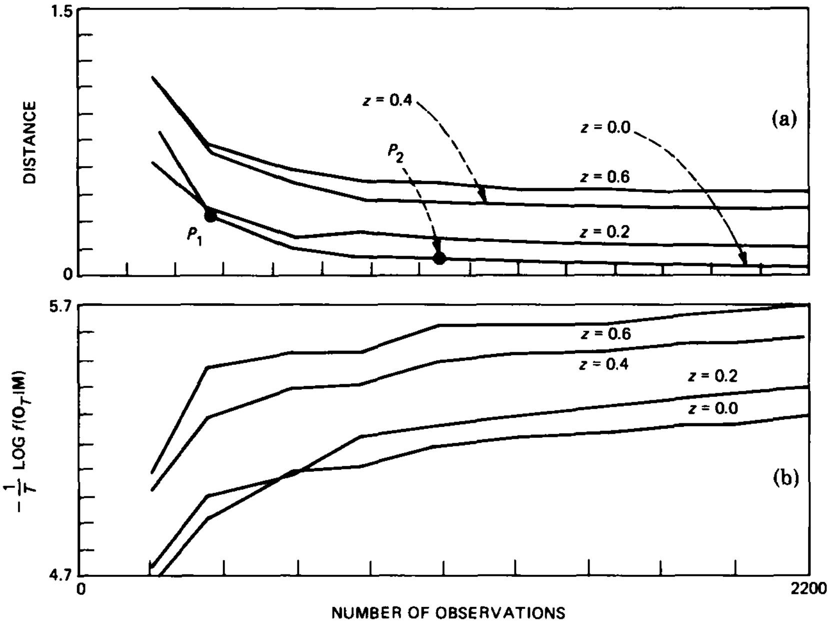

Fig. 3—(a) Distance and (b) average log likelihood as a function of the number of observations in the training sequence, and as a function of the initial estimation deviation, z.

estimated model. When a smaller number of observations is generated for estimation, the statistical characteristics of the source become less well represented in the generated observation sequences, and hence, the estimated model is more data specific, resulting in greater variations in the log likelihood. This can be more easily seen from the behavior of distance for a specific set of training sequences as the reestimation procedure iterates to a stable solution. Such a plot is given in Fig. 4 for $Q = 20$ sequences (part a) and for $Q = 50$ sequences (part b). (Note that these two curves show the distance behavior of the model reestimate as it converges to the solution corresponding to the two particular points, P_1 and P_2, in the upper curve of Fig. 3, respectively.) For $Q = 20$ sequences, the training set does not provide a good characterization of the source model—it is too short; hence the model distance decreases for a couple of iterations and then increases as the local estimated parameters are adjusted to match those of the specific observation sequence rather than those of the true generating model. For $Q = 50$ sequences, the distance between the estimated model and the true source model steadily decreases.

3.3 Sensitivity of model reestimation to evaluation of the density function

Because of the wide dynamic range of the density function, $b_j(\mathbf{x})$, of

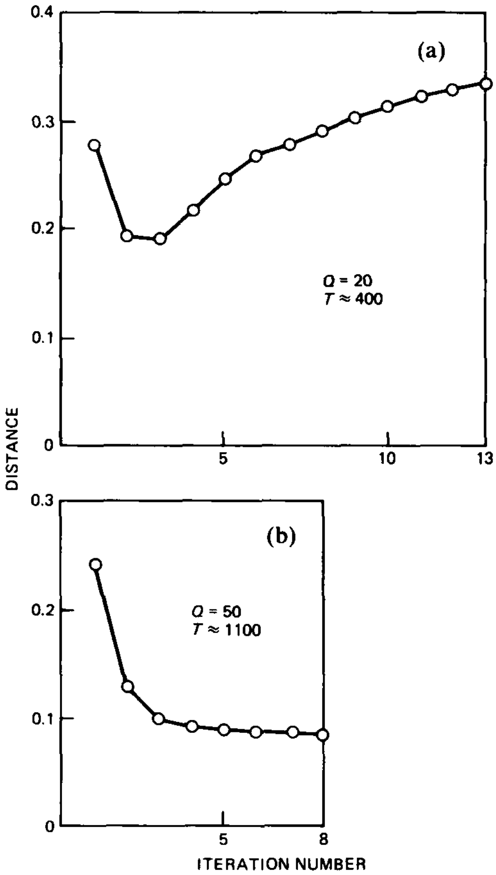

Fig. 4—(a) Distance for $Q = 20$ sequences training and (b) $Q = 50$ sequences training as a function of the iteration number.

eq. (1)—especially when the estimates of μ and $\mathbf{U}$ are in error—a minimum value clipping level, f_{CLIP}, is usually required to avoid potential underflow and singularity problems. In our study, whenever $b_j(\mathbf{x})$ was less than $10^{-f_{\text{CLIP}}}$, it was artificially clamped at $10^{-f_{\text{CLIP}}}$; otherwise $b_j(\mathbf{x})$ was kept as computed. This, in effect, injects certain noise components into the observations. To understand the effect of f_{CLIP} on the resulting model estimates, a left-to-right, four-state, one-mixture, two-dimensional model was used, with the specification

$$N = 4, \quad M = 1, \quad K = 2$$

$$\mathbf{A} = \begin{bmatrix} .8 & .15 & .05 & 0 \\ 0 & .8 & .15 & .05 \\ 0 & 0 & .8 & .2 \\ 0 & 0 & 0 & .1 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\mu_{1.} = [0 \quad 0], \qquad \mu_{2.} = [4 \quad 4], \qquad \mu_{3.} = [8 \quad 8], \qquad \mu_{4.} = [12 \quad 12]$$

$$\mathbf{U}_{j1,k,l} = \begin{cases} 1, & k = l, \quad \text{all} \quad j \\ .1, & k \neq l, \quad \text{all} \quad j. \end{cases}$$

The value of f_{CLIP} was varied from 70—which is essentially full precision on the Data General MV8000 32-bit computer—down to 10—severe clipping—and initial estimates of μ' were generated as in eq. (10). For each value of f_{CLIP} and z, the distance between the model resulting from the reestimation procedure and the original source model was computed, and the results are plotted in Fig. 5. For all runs, the number of observation sequences used in training was 50; hence there was an adequate number of observations for the parameter estimates to converge to the true model. The results given in Fig. 5 show that for $z = 0$ the distance is insensitive to values of f_{CLIP} over the entire range. For $z = 0.2$, the distance is much larger for $f_{\text{CLIP}} = 10$ than for all other values of f_{CLIP}. The differences in distance between the results for $z = 0.2$ and those for $z = 0.0$ are insignificant except for those at $f_{\text{CLIP}} = 10$. For $z = 0.4$, the model estimates yield larger distances than for $z = 0$ or $z = 0.2$ for all values of f_{CLIP}. The differences for f_{CLIP} values of less than 50 are primarily due to the sensitivity of the reestimation procedure to the initial μ estimates as discussed previously. The differences for f_{CLIP} in the range 10 to 40 are due to sensitivities of the reestimation procedure to the clipping itself. To understand this sensitivity, consider Fig. 6, which shows a Gaussian with a clipping threshold $10^{-f_{\text{CLIP}}}$. In the case that initial estimates of μ (and $\mathbf{U}$) are very close to the true value, the density function will rarely, if ever, be clipped; hence until $10^{-f_{\text{CLIP}}}$ approaches the peak of the density function, the clipping has little effect on the model estimate. In the case where initial estimates of μ are far from the true value, a large percentage of the density computations will be clipped and the reestimation procedure will be unable to improve the parameter estimates because the density function is essentially flat in the region of the clipping. For such cases, very poor estimates of μ result and large model distances are obtained. The results point out an important consideration in practical implementations of the estimation algorithm, where finite precision is inevitable.

3.4 Modeling correlated processes by mixtures of uncorrelated processes

The mixture form of eq. (1) is a very versatile and flexible representation of the pdf in each state. For example, a complicated multivariate pdf may be approximated by a mixture of Gaussian multivariate densities with full covariance matrices, or, by increasing the number of mixture components, a mixture of Gaussian multivariate densities with only diagonal covariance matrices (i.e., vector elements are un-

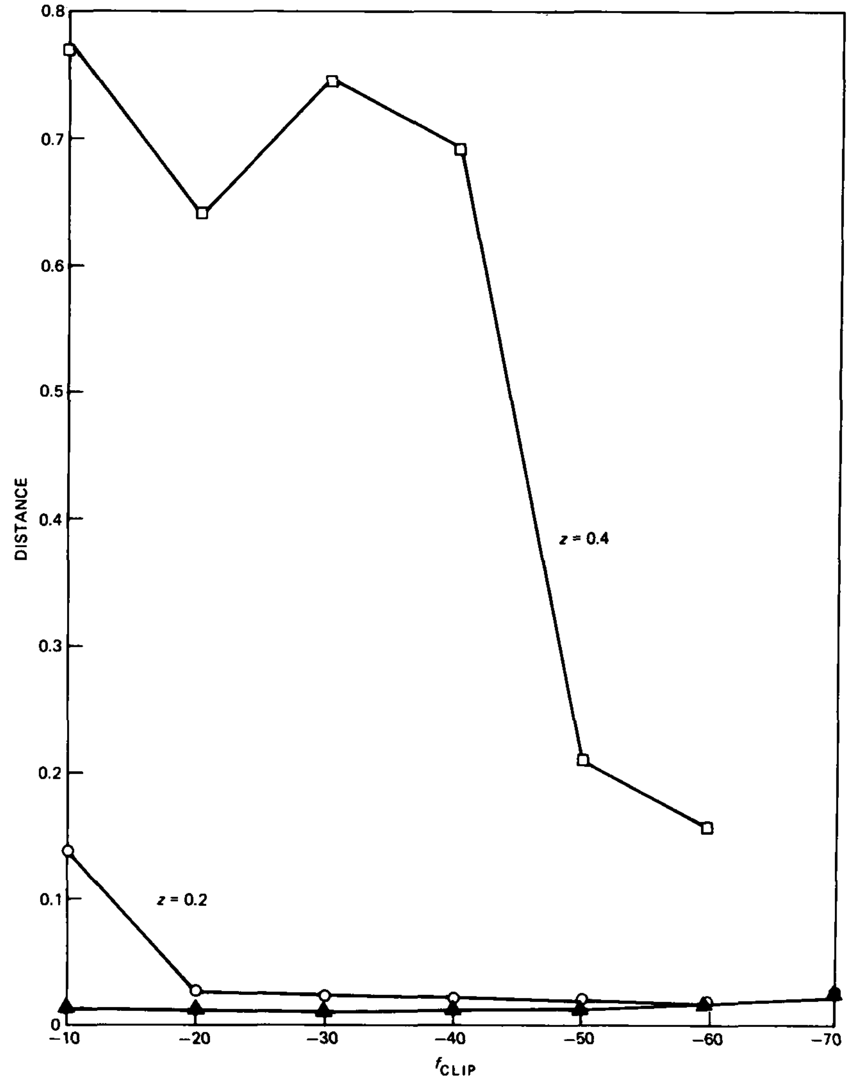

Fig. 5—Distance as a function of the density clipping threshold for $z = 0$, 0.2, and 0.4.

correlated). To better understand this concept, we studied the tradeoff between the degree of correlation and the number of mixture components in the representation by modeling correlated multivariate densities with different numbers of uncorrelated multivariate densities using the HMM framework.

The source model used for these studies had the following specifications:

$$N = 4, \qquad M = 1, \qquad K = 2$$

$$\mathbf{A} = \begin{bmatrix} .8 & .15 & .05 & 0 \\ 0 & .8 & .15 & .05 \\ 0 & 0 & .8 & .2 \\ 0 & 0 & 0 & 1.0 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\boldsymbol{\mu}_{1.} = [0 \quad 0], \qquad \boldsymbol{\mu}_{2.} = [4 \quad 4], \qquad \boldsymbol{\mu}_{3.} = [8 \quad 8], \qquad \boldsymbol{\mu}_{4.} = [12 \quad 12]$$

$$\mathbf{U}_{j1kl} = \begin{cases} 1, & k = l \\ \rho, & k \neq l, \end{cases}$$

where ρ varied from 0 to 0.9 (in steps of 0.1). Thus, the source model had a full two-dimensional covariance matrix with correlation ρ between components of each vector.

We considered two separate HMMs for matching the observation sequences of the full covariance source model. The first model used an $M = 1$ (a single) mixture with a diagonal covariance matrix; the second model used an $M = 5$ mixture, where each component density again had a diagonal covariance matrix. Since we were interested only in the capabilities of the models—and not in the concomitant problems

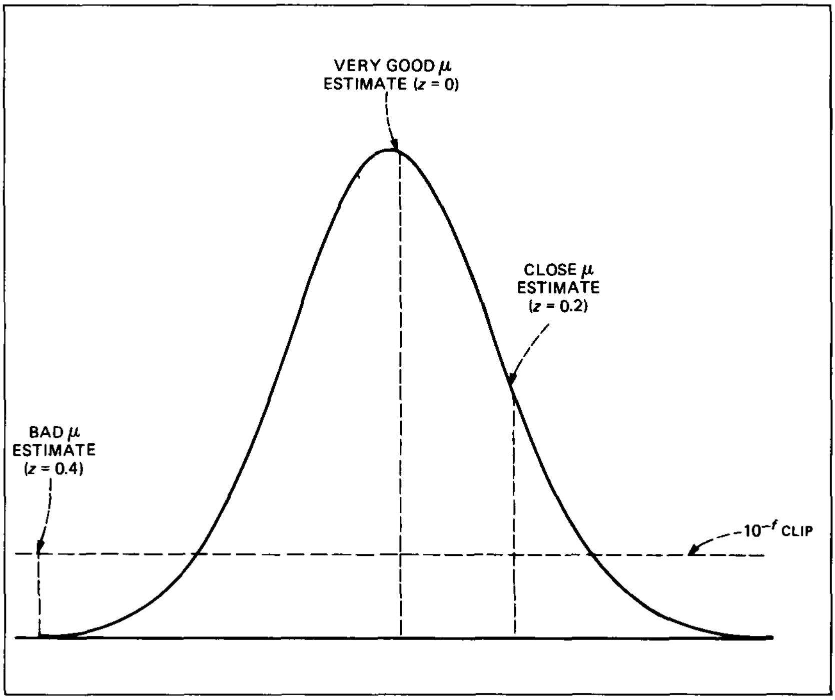
VERY GOOD μ ESTIMATE (z = 0)
CLOSE μ ESTIMATE (z = 0.2)
BAD μ ESTIMATE (z = 0.4)
10^{-f} CLIP

Fig. 6—Explanation of the sensitivity of the reestimation algorithm to the density clipping threshold for different values of z.

of reestimation—the initial estimates of the model parameters were selected to optimize the match. Thus for the $M = 1$ model, the initial estimates were identical to those of the source, except the off-diagonal covariance terms were set to 0. For the $M = 5$ model, the initial estimates of $\boldsymbol{\mu}$ and $\mathbf{U}$ were adjusted in order to best match the full covariance with correlation ρ by the $M = 5$ mixtures. The procedure used is illustrated in Fig. 7, which shows a K equals a two-dimensional correlation in the (x_1, x_2) plane (part a), and a one-dimensional slice (part b). The initial estimates of $\boldsymbol{\mu}$ for the $M = 5$ case are shown by the center dots of the five circles in part a. The mixture gains and the mixture covariances were chosen to provide good initial fits to the correlated covariance as shown in both parts a and b.

The results of estimating optimum models for the $M = 1$ and $M = 5$ cases are shown in Fig. 8, which gives plots of model distance versus ρ. For $M = 1$, the model fits have distances less than or equal to 0.1 only for $\rho \leq 0.35$, and have distances less than or equal to 0.2 for $\rho \leq 0.5$. For $M = 5$, the model fits have distances less than or equal to 0.1 for $\rho \leq 0.7$, and have distances less than or equal to 0.2 for ρ up to 0.9. Thus, for this case, the $M = 5$ mixture models without correlations provide excellent approximations to models with correlated random variables up to correlations of 0.9.

The results presented above show that it is possible to model a K-dimensional ($K = 2$) correlated random process by a mixture of M-uncorrelated, K-dimensional, Gaussian random processes. The question that remains is why one would be interested in using such an approximation. There are two possible reasons that readily come to mind. First, there is the possibility that more reliable estimates can be made of the set of $2M \cdot K$ means and variances for the M-mixture uncorrelated processes case, than for the set of $K(K + 3)/2$ means and correlations for the one-mixture correlated process. If this is the case the trade-off is between the increased error in the approximation process and the increased reliability in the estimation process. The second possible reason for using the M-mixture uncorrelated process instead of the one-mixture correlated process is the potential for a decrease in the number of parameters that need to be estimated. To see when this can occur, we define P_c as the number of parameters for the one-mixture correlated density, and P_μ as the number of parameters in the M-mixture uncorrelated density case. Then, assuming K-dimensional vectors, we get

$$P_c = \frac{K(K + 3)}{2}$$

and

$$P_\mu = M(2K + 1).$$

For $P_\mu \leq P_c$ we require

$$M \leq \frac{K(K+3)}{2(2K+1)}.$$

Thus, for $K \leq 5$, the largest M can be is 1, for $9 \geq K \geq 6$, the largest M can be is 2, and for $13 \geq K \geq 10$, the largest M can be is 3 to realize

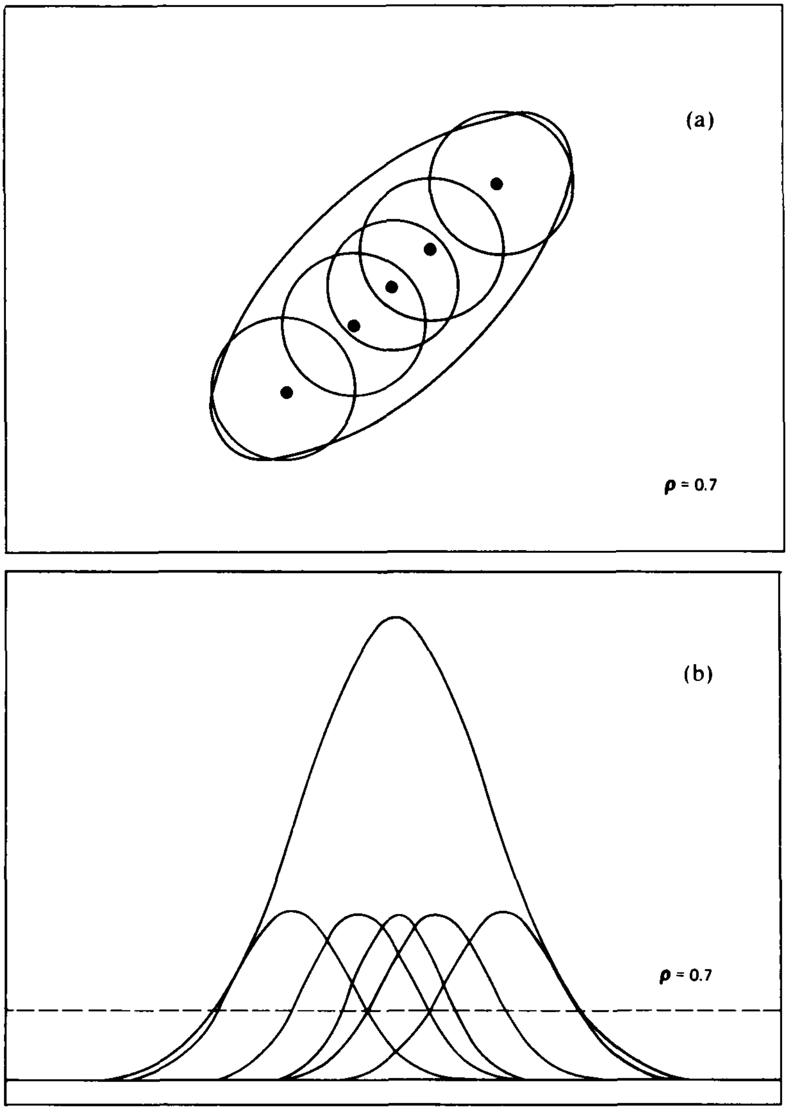

Fig. 7—(a) Observation region in the (x_1, x_2) plane for highly correlated vector components along with initial estimates of μ for $M = 5$ model; (b) interpretation of initial estimates along a one-dimensional projection of the (x_1, x_2) plane.

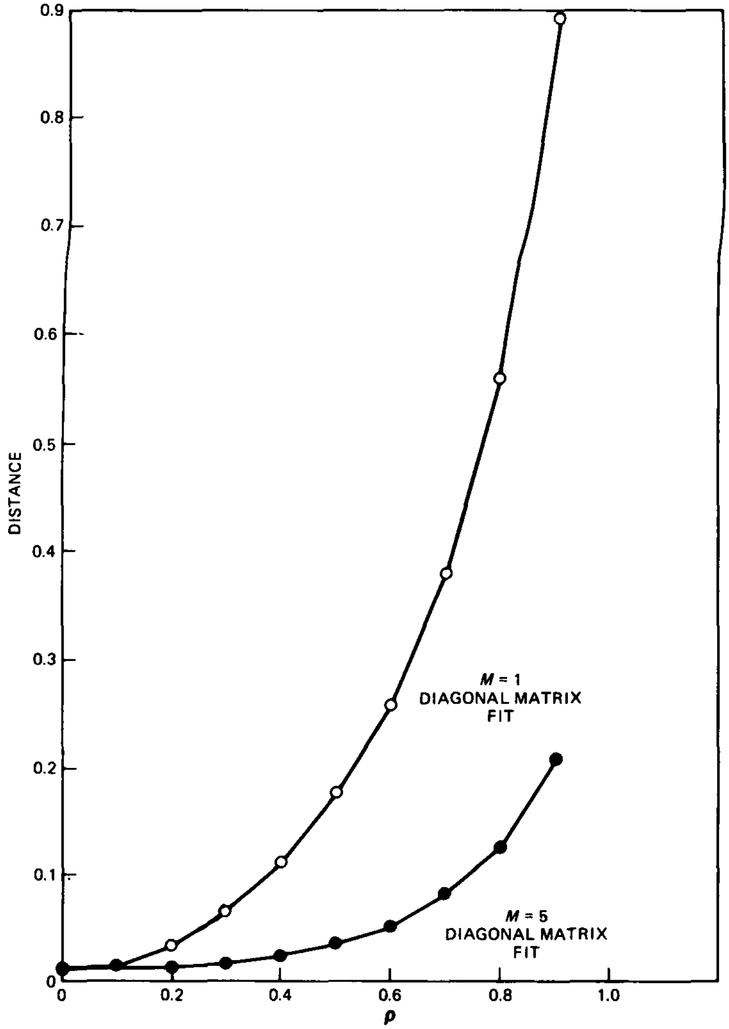

Fig. 8—Distance versus ρ for $M = 1$ and $M = 5$ mixture fits using diagonal covariance matrices.

any reduction in the number of variables to be estimated. For speech-recognition applications, we generally use $K \approx 10$; hence values of $M \leq 3$ could be considered. Whether or not the model is adequately represented with this many diagonal mixtures depends heavily on the specific application. The purpose of the above discussion is to point out the possibilities of the alternative method.

IV. DISCUSSION

The results presented in the previous section have shown the following:

1. Continuous HMMs characterized by mixture densities are most sensitive to estimation errors in the locations of the means of each mixture density. If the error in the initial estimate of the mean becomes sufficiently large, then the reestimation procedure has very little chance of giving good model parameter estimates.
2. The sensitivity of the models to errors in initial covariance estimates is less than that due to errors in the initial mean estimates.
3. The sensitivity of the models to errors in either transition matrix coefficients, or mixture gains, is low. Hence, good model estimates can be obtained even with poor initial estimates of these parameters, as long as the distribution does not contain singularities.
4. We have found that observations on the order of 500 to 1000 are adequate for models that are typical of many applications in speech processing (e.g., models with $N = 10$, $K = 10$, $M = 3$).
5. Good initial parameter estimates become critical in the reestimation procedure when word precision for the evaluation of the density function is limited—an inevitable situation in practical implementations.
6. Mixture density models with diagonal covariance matrices for each mixture can be used to approximate full covariance models.

The most important conclusion from our experiments is that it is absolutely mandatory to have a good initial guess of the means of the density functions to obtain good HMMs. With a good initial guess of the means, the parameter reestimation procedure is capable of yielding good models even if other model parameters have poor intial estimates.

V. SUMMARY

Several interesting properties of continuous density HMMs have been discussed. These include model sensitivity to initial parameter estimates, to evaluation of the density function, and to size and type of training sequence. We have shown how a mixture density of uncorrelated variables can successfully represent a model with highly correlated variables, as long as enough mixtures are used. The results presented here can be applied to a variety of real-world problems.

REFERENCES

1. R. L. Cave and L. P. Neuwirth, "Hidden Markov Models for English," *Hidden Markov Models for Speech*, J. Ferguson, ed., IDA-CRD, 1980, pp. 16–56.
2. L. R. Bahl et al., "Optimizing Decoding of Linear Codes for Minimizing Symbol Error Rate," IEEE Trans. on Inform. Theory, *IT-20*, No. 2 (March 1984), pp. 284–7.
3. L. E. Baum and J. A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology," Bull. Amer. Math. Soc., *73* (1967), pp. 360–3.
4. J. K. Baker, "The Dragon System—An Overview," IEEE Trans. on Acoust., Speech, Signal Processing, *ASSP-23*, No. 1 (February 1975), pp. 24–9.

5. L. R. Bahl, F. Jellinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Trans. Pattern Analysis Mach. Intell., *PAM 1-5,* No. 2 (March 1983), pp. 179–90.
6. L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated Word Recognition," B.S.T.J., *62,* No. 4 (April 1983), pp. 1075–106.
7. L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," Ann. Math. Statist., *41* (1970), pp. 164–71.
8. L. R. Liporace, "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," IEEE Trans. Inform. Theory, *IT-28* (September 1982), pp. 729–34.
9. B. H. Juang, S. E. Levinson, and M. M. Sondhi, unpublished work.
10. L. E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process," Inequalities, *3* (1977), pp. 1–8.
11. B. H. Juang and L. R. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models," *64,* No. 2 (February 1985), pp. 391–408.

AUTHORS

Biing-Hwang Juang, B.Sc. (Electrical Engineering), 1973, National Taiwan University, Republic of China; M.Sc. and Ph.D. (Electrical and Computer Engineering), University of California, Santa Barbara, 1979 and 1981, respectively; Speech Communications Research Laboratory (SCRL), 1978; Signal Technology, Inc., 1979–1982; AT&T Bell Laboratories, 1982; AT&T Information Systems Laboratories, 1983; AT&T Bell Laboratories, 1983—. Before joining AT&T Bell Laboratories, Mr. Juang worked on vocal tract modeling at the Speech Communications Research Laboratory, and on speech coding and interference suppression at Signal Technology, Inc. Presently, he is a member of the Acoustics Research Department, where he is researching speech communications techniques and stochastic modeling of speech signals.

Stephen E. Levinson, B.A. (Engineering Sciences), 1966, Harvard; M.S. and Ph.D. (Electrical Engineering), University of Rhode Island, Kingston, Rhode Island, 1972 and 1974, respectively; General Dynamics, 1966–1969; Yale University, 1974-1976; AT&T Bell Laboratories, 1976—. From 1966 to 1969, Mr. Levinson was a design engineer at Electric Boat Division of General Dynamics in Groton, Connecticut. From 1974 to 1976, he held a J. Willard Gibbs Instructorship in Computer Science at Yale University. In 1976, he joined the technical staff at AT&T Bell Laboratories, where is is pursuing research in the areas of speech recognition and cybernetics. Member, Association for Computing Machinery, Acoustical Society of America, editorial board of Speech Technology; associate editor, IEEE Transactions on Acoustics, Speech and Signal Processing.

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D., 1967 (Electrical Engineering), The Massachusetts Institute of Technology; AT&T Bell Laboratories, 1962—. Presently, Mr. Rabiner is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice-Hall, 1978), and *Multirate Digital Signal Processing* (Prentice-Hall, 1983). Member, National Academy of Engineering, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America, IEEE.

Man Mohan Sondhi, B.Sc. (Physics), Honours degree, 1950, Delhi University, Delhi, India; D.I.I.Sc. (Communications Engineering), 1953, Indian Institute of Science, Bangalore, India; M.S., 1955 and Ph.D. (Electrical Engineering), 1957, University of Wisconsin; AT&T Bell Laboratories, 1962—. Before joining AT&T Bell Laboratories, Mr. Sondhi worked for a year at the Central Electronics Engineering Research Institute, Pilani, India, and taught for a year at the University of Toronto. At AT&T Bell Laboratories his research has included work on speech signal processing, echo cancellation, adaptive filtering, modeling of auditory and visual processes, and acoustical inverse problems. From 1971 to 1972, Mr. Sondhi was a guest scientist at the Royal Institute of Technology, Stockholm, Sweden.